11-2018

Multi-modal learning using deep neural networks

Dheeraj Kumar Peri
dp1248@rit.edu

Multi-modal learning using deep neural networks

by

Dheeraj Kumar Peri

November 2018

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Engineering
Department of Computer Engineering
Kate Gleason College of Engineering
Rochester Institute of Technology

Approved By:

Date:_____

Dr. Raymond Ptucha
Primary Advisor – R.I.T. Dept. of Computer Engineering

Date:_____

Dr. Andreas Savakis
Committee Member – R.I.T. Dept. of Computer Engineering

Date:_____

Dr. Andres Kwasinski
Committee Member – R.I.T. Dept. of Computer Engineering

Acknowledgements

This journey of research and the constant hunt for improving things by learning ways to analyze and implement new algorithms has left a profound impact on me. I am immensely grateful to my advisor, Dr. Raymond Ptucha for getting me interested in deep learning research and for always pushing me to do better. The team meetings and hours of brainstorming new ideas with him has always been amazing. I would also like to thank members of the Machine Intelligence Lab for making the lab a friendly environment and a fun place to work. I would like to thank my father, mother and my brother for always supporting me and encouraging me to pursue my interests.

ABSTRACT

Humans have an incredible ability to process and understand information from multiple sources such as images, video, text, and speech. Recent success of deep neural networks has enabled us to develop algorithms which give machines the ability to understand and interpret this information. Convolutional Neural Networks (CNN) have become a standard in extracting rich features from visual stimuli. Recurrent Neural Networks (RNNs) and its variants such as Long Short Term Memory (LSTMs) units have been highly successful in encoding and decoding sequential information like speech and text. Although these networks are highly successful when applied to narrow applications, there is a need to both broaden their applicability and develop methods which correlate visual information along with semantic content.

This master's thesis develops a common vector space between images and text. This vector space maps similar concepts, such as pictures of dogs and the word "puppy" close, while mapping disparate concepts far apart. Most cross-modal problems are solved using deep neural networks trained for specific tasks. This research formulates a unified model using CNN and RNN which projects images and text into a common embedding space and also decodes the image and text embeddings into meaningful sentences. This model shows diverse applications in cross modal retrieval, image captioning and sentence paraphrasing and shows promising directions for neural networks to generalize well on different tasks.

Table of Contents

List of Figures

List of Tables

Acronyms

CNN

Convolutional Neural network

CVS

Common Vector Space

GRU

Gated Recurrent Unit

LSTM

Long Short Term Memory unit

RNN

Recurrent Neural network

STT

Show, Translate and Tell

STT-ATT

Show, Translate and Tell with Attention

Chapter 1

Introduction

1.1 Introduction

One of the long standing goals of artificial intelligence is for machines to learn and understand the dynamics of complex environments. Infants learn to perform tasks and gain skills by interacting with the environment through visual and language information. Deep learning has enabled machines to understand such complex interactions and generalize well to new scenarios. Machine learning algorithms train on huge amount of data from images, text, audio, video etc. and try to come up with a function that closely represents the mapping between input and the desired output. For example, in an image classification problem, the algorithms learn to correlate the information in the image of a cat to the label 'cat' by continuously updating their internal parameters. This automated learning of features has replaced the use of traditional methods like Histogram of Oriented Gradients (HoG) [39] and Scale Invariant Features [40]. Recent success of Convolutional Neural Networks (CNN) for encoding images and Recurrent Neural Networks (RNN) for representing text information can be attributed to the back-propogation algorithm [41] which stochastically updates model parameters and guides the learning process. This work attempts to develop a Common Vector Space (CVS) which embeds both images and text. Similar concepts such as an image of a dog and the descriptions related to a dog are mapped close while dissimilar concepts are mapped far apart. A unified model is developed which can generalize well over different cross modal applications.

1.2 Contributions

The main contributions of this thesis work can be summarized as follows

- A unified model which jointly trains on images and captions and learns to generate new captions given either an image or a text as a query.
- Diverse applications of the joint model on three different tasks, namely image captioning, cross modal retrieval and sentence paraphrasing.

1.3 Background

1.3.1 Convolutional Neural Network:

Convolutional neural networks have become the defacto-standard for the tasks of image classification, segmentation and detection. Typically they comprise of the following layers:

- Convolution layer
- Pooling layer
- Activation layer
- Fully connected layer

A convolution layer consists of multiple filters which slide across the input image and produce a linear response from filter weights applied to the input pixels. Each filter learns a different set of representations of the original image eg: color, shape and edge information.

A pooling layer aggregates the information across a specified window in an image. The two popular pooling approaches used are max pooling and average pooling. Max pooling outputs the maximum of the pixels in the window under consideration whereas average pooling

outputs the average intensity of the pixels. Performing pooling reduces the spatial dimensions of the input.

An activation layer introduces non-linearity in the network. It helps to learn complex representations that exist between input image and desired target in the network.

A fully connected layer is used as a final layer in most of the classification problems. It is generally used to transform a high dimensional representation into an n-dimensional representation by connecting all the pixels in the input layer to each neuron in the output layer.

During training, the filters of the convolutional layers and weights of fully connected layers are learned by optimizing the cross entropy loss between predicted and groundtruth labels of samples in classification problems. An example of a typical CNN is shown in Figure 1.

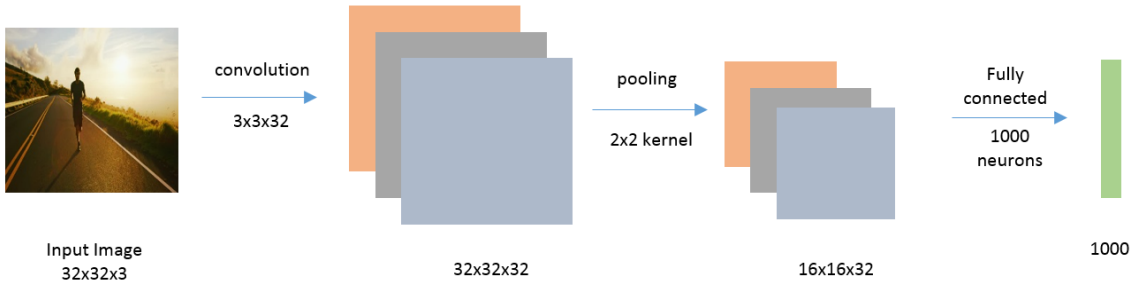


Figure 1 An example Convolutional Neural Network.

1.3.2 Recurrent Neural Networks

Recurrent Neural Networks (RNN) have achieved significant success in time-series problems and machine translation. A basic RNN unit consists of a hidden state and an input which together predict the next state of a sequence. Some of the most popular variants of RNNs are Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) units. LSTMs are often the preferred choice for long sequences as they tend to remember long term dependencies by using gating mechanisms. Figure 2 shows an example of a single LSTM unit

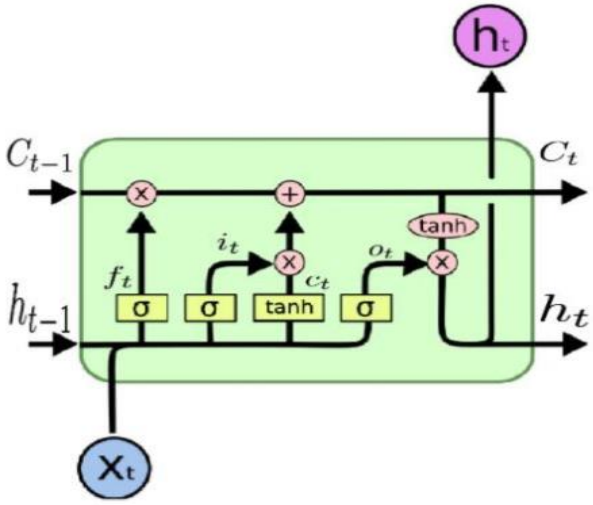


Figure 2 Basic LSTM cell [42].

where c_t denotes the memory unit, h_t denotes the hidden state, f_t denotes the forget gate, i_t denotes input gate and o_t denotes the output gate.

The above gates followed by sigmoid and tanh activation units regulate the amount of information that needs to be passed to the consecutive time steps in the network. More commonly, LSTM networks are used in machine translation, which are otherwise known as sequence-sequnce models.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \quad (1)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1}\right)$$

$$g_t = \emptyset\left(W_{xc}x_t + W_{hf}h_{t-1}\right)$$

$$c_t = f_t c_{t-1} + i_t g_t$$

$$h_t = o_t \emptyset(c_t)$$

where i_t, o_t, f_t and g_t are the input gate, output gate, forget gate and input node respectively. The cell memory state is given by c_t which contains the overall information about

the cell. The hidden state h_t is passed to future timesteps in the network which contains the aggregate information of the previous timesteps.

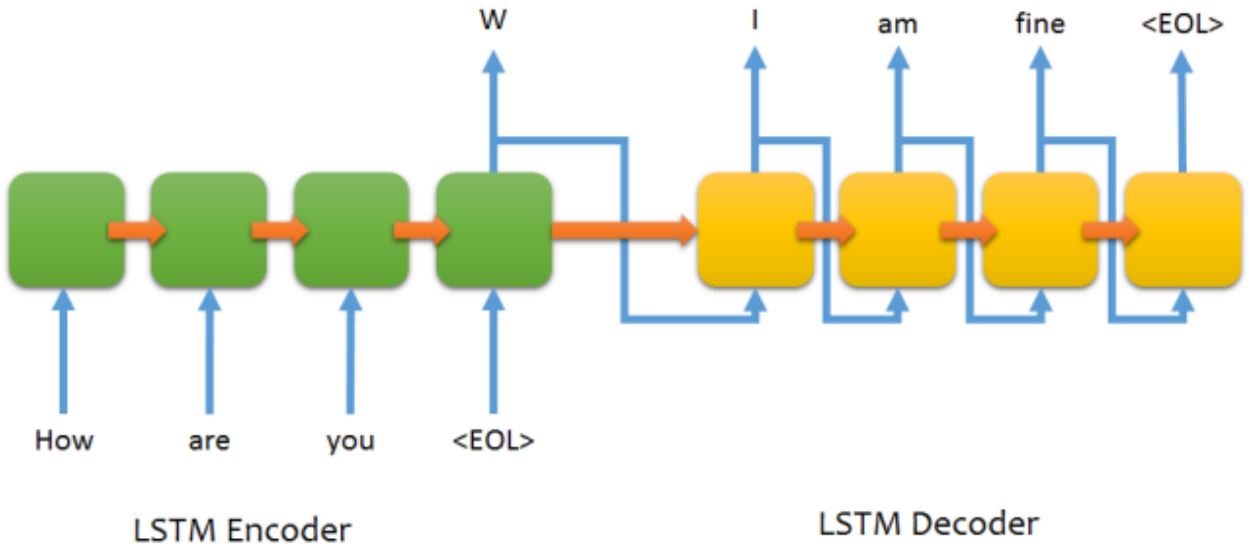


Figure 3 Long Short Term Memory network with encoder and decoder chains [6].

Figure 3 shows an example of an encoder-decoder network with LSTM units. The encoder and decoder may or may not share the same LSTM units. The encoder encodes the input sequence "How are you", one word at a time using a word embedding. The final state of the encoder is the last hidden timestep of the input sequence. The encoder's final time step is passed along with a start token as input to the first timestep of the decoder. The decoder is unrolled for variable timesteps and outputs a decoded sentence followed by an end token. This framework has shown promising results in machine translation and sentence paraphrasing.

Chapter 2

Related Work

2.1 Cross Modal Applications

Image captioning was one of the earliest works that demonstrated outstanding capabilities of neural networks to generalize well on learning patterns in both vision and language modalities. Neural networks trained with back propogation tend to learn patterns in the image and correlate the relationships between objects in the image and individual words in the sentence. The branch of study that deals with similarities between different entities is called Metric Learning. The task of cross-modal retrieval involves learning similar representations between two modalities. For example, given the two modalities of image and text, one can extract meaningful content from a database given a query of either modality. Images have diverse content and a sentence describing the image should capture not only the objects present in the image but also the relationship between them. Often images can be described in many ways and capturing the right context in the sentence is challenging. For example, "a man is running" and "a man is not running" have most of words same but the word "not" changes the entire meaning of the description. CNNs have become the defacto standard in representing images and recurrent neural networks have been adept at capturing the syntactic and semantic representations of the sentence. In this thesis, neural networks with latest CNN and RNN architectures and current metric learning approaches are explored in cross-modal settings to enhance image2text, image2image, text2image, and text2text retrieval.

2.2 Current Metric Learning Approaches

Metric learning models involving images and text include:

- Extracting features from images and text using CNNs and language models (Bag of Words, LSTMs, skipgram)
- Generate embeddings from these features using fully connected layers.
- Form positive and negative pairing of data and use different loss functions for convergence.

Some of the commonly used loss functions are:

1) Contrastive Loss

In contrastive learning, positive and negative pairs of the data are formed by the distance between the image and caption encoding. Contrastive loss strives to have negative pairs be at least a *margin* distance away from positive pairs. The loss function is as follows:

$$L_c = \frac{1}{2N}\sum\big((y)d^2 + (1-y)\max(margin - d, 0)^2\big) \quad (2)$$

where *d* is the distance between the vectors in a pair. The first term minimizes the distance between positive pairs, while the second term penalizes negative samples whose distance is closer than a *margin*. The distance can be Euclidean, cosine, or other appropriate metric.

2) Triplet Loss

Triplets are formed by selecting an anchor sample and generating positive and negative examples with respect to the anchor sample. The distance between the positive sample and the anchor is minimized whereas the distance between the anchor and negative sample is maximized.

$$L_c = \frac{1}{2N}\sum_{i=0}^{N} \max(0,\ \left|f_a^i - f_p^i\right|^2 - \left|f_a^i - f_n^i\right|^2 + \ margin) \quad (3)$$

where f_a^i is the feature embedding of the anchor, f_p^i is the feature embedding of positive sample and f_n^i is the feature embedding of negative sample. The triplet learning process is shown in the Figure 4.

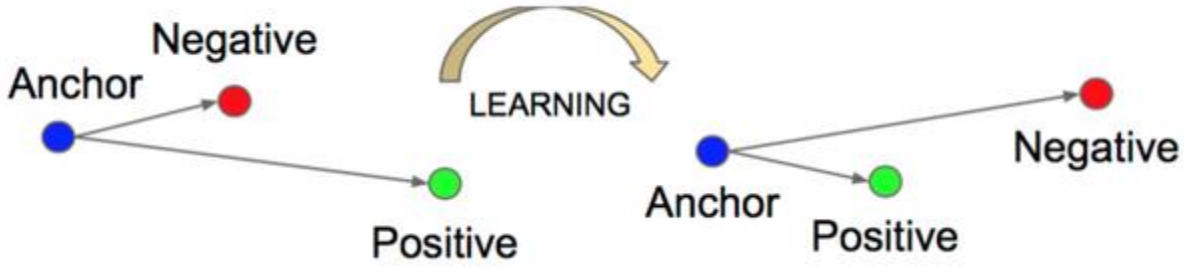


Figure 4 Optimizing latent space through triplet loss [12].

3) Lifted Structured Loss

Lifted structure loss extends the concept of triplet loss by considering multiple negative samples for each positive sample. It ensures the distance between the positive and anchor sample is less than distance between anchor and all other negative samples in the batch. The lifted loss is defined as follows

$$L_c = \frac{1}{2|P|}\sum_{(i,j)\in P} \max(0, J_{i,j}) \quad (4)$$

where

$$J_{i,j} = \max\left(\max_{(i,k)\in N} \alpha - D_{i,k}\ ,\ \max_{(j,l)\in N} \alpha - D_{j,l}\right) \quad (5)$$

where N denotes the set of negative samples and P denotes the set of positive samples. Each sample is compared against a positive sample and all other negative samples in the batch thereby forming tighter boundaries between samples during training.

Many approaches either use the above losses or the extensions of these to optimize the distance between embeddings of different modalities. Most of the metric learning approaches use labels as anchors to form positive and negative pairing. Two pictures of dogs are treated as similar examples irrespective of their color, orientation and background, whereas a picture of dog and cat are treated as negative pair.

On the contrary, in cross modal setting, there might not be labelled data with exclusive categories for each image and captions. More naturally occurring images and text contain multiple objects and various kind of actions describing their context. This poses a harder challenge to distinguish samples of any specific category. The general consensus is that the captions that were used to describe the image are treated as positive pairs and the rest of the captions in a dataset are treated as negative pairs. This assumption makes the general metric learning loss functions applicable to the problem of cross-modal retrieval.

2.3 Related Work

Koch *et al.* [17] introduced siamese networks to learn similarities between characters using contrastive loss and achieved superior performance on one-shot image recognition tasks. Jiquan *et al.* [1] demonstrated that better features can be learned if multiple modalities are present during training. They also demonstrate a method to learn shared representations of different modalities. Scott *et al.* [2] proposed Deep Structured Joint Embedding (DSJE) which includes joint training of images and text and they show improved results on retrieval and zero-

shot recognition tasks. Schroff *et al.* [12] proposed triplet loss to enhance similatity learning by considering triplets of data and showed improved performance on face recognition. Hoffer *et al*. [18] used triplet loss on metric learning problems and compared its performance to siamese networks which used contrastive loss. Song *et al*. [15] proposed lifted structured loss which essentially takes advantage of all the samples in the batch. For each positive sample, it pushes all the negative samples away by a margin in a batch. This showed improved retrieval performance on standard benchmark datasets. Euclidean distance is used as standard distance metric in their experiments.

One of the most important aspect in similarity learning is the distribution of samples in a batch and the strategy of forming positive and negative pairs within a batch. Not all the negative samples are equally negative. During training, optimizing Euclidean loss of positive and negative samples with respect to an anchor sample in a batch results in the formation of discrete clusters in the high dimensional space. Negative samples can be classified into three categories as follows

- Hard negatives
- Semi-hard negatives
- Easy negatives

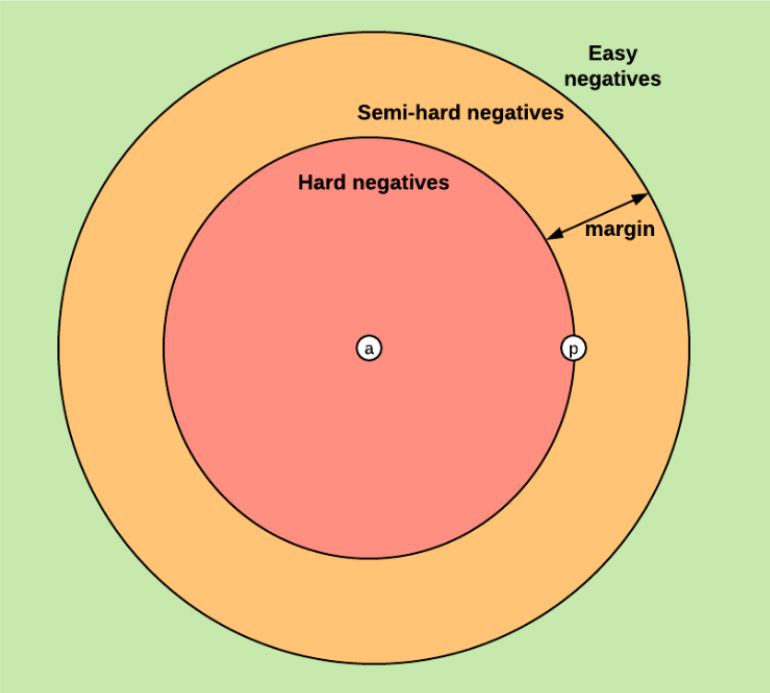


Figure 5 Distrubution of negative samples [20].

Figure 5 shows the distrinution of negative samples with respect to an anchor sample. Hard negative samples are closer to positive samples, semi-hard negatives lie within margin distance from the positive samples and easy negatives already are far away from the anchor sample under consideration. Hard negative mining is a strategy that mines the hardest negatives for a given sample in a batch. Although hard negatives produce a high loss value, they also produce high gradients which might lead to bad convergence of the model.

Exploiting the success of generative adversarial networks [16], Duan *et al.* [19] proposed to use a generator that exploits all easy negative samples and transforms them into hard negative samples. A generator is trained adversarially to generate features which are similar to features from hard negative samples thereby enhancing the training. A combination of adversarial loss as well as metric learning loss functions helped in exploiting more

discriminative features from the network and improved retrieval results compared to standard metric loss functions. Figure 6 shows their model where the generator is a three layered fully connected network which generates synthetic negative samples.

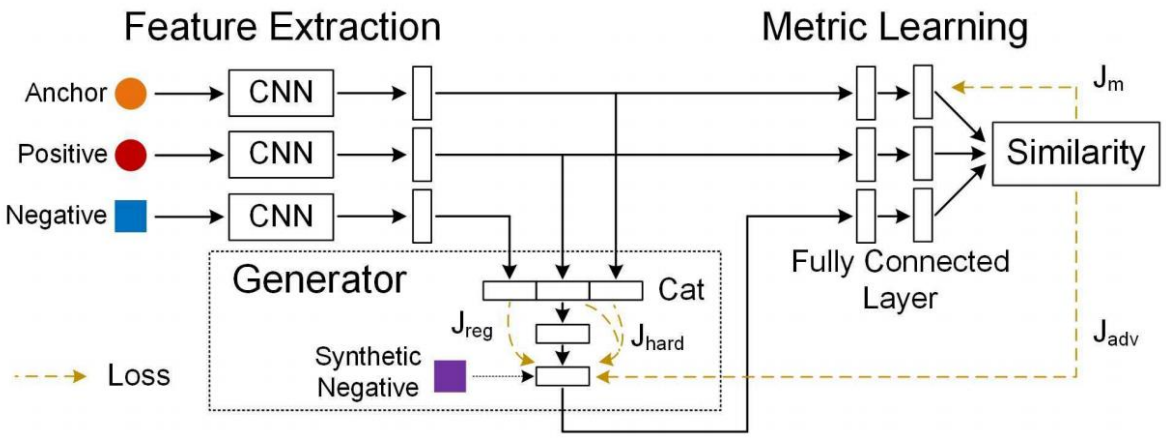


Figure 6 Deep Adversarial Metric Learning[19].

Schroff *et al.* [12] proposed to use semi-hard negative mining which samples only semi-hard negatives for each sample. They found loss to be decaying smoothly compared to random negative sampling. Chao *et al.* [21] proposed a margin based loss and also proposed distance weighted sampling which selects negative samples based on their distances. They show that learning the margin parameter removes the inherent bias that restricts all negative samples to be pushed apart by a constant margin value.

2.4. Image-Text models

The main difference between metric learning and multi-modal learning is the encoding of different modalities in the shared high dimensional space. The similarity between images

can be expressed in the form of Euclidean distance between vector representations of these images. This vectorization of images is generally a vector representation from a fully connected layer of a CNN. The effectiveness of the CNN also plays an important factor in learning discriminative features among images. In the context of multi-modal learning which involves images and captions, separate encoders for each modality are required due to difference in the structures. Images are encoded by CNN (image2vec) and a caption is passed through an RNN (sent2vec). Aviv *et al*. [22] used two way neural networks to optimize Euclidean loss between images and text in a common embedding space. Vendrov *et al*. [23] proposed to use order-violation penalty to enforce constraint on the order in which the embeddings are learned. In particular, they only use absolute value of image and text embeddings and use margin-based loss to optimize the model.

Faghri *et al.* [23] proved hard negative mining can be useful and they showed significant improvements on cross modal retrieval problems. This is counterintuitive to metric learning problems where hard negative mining hurts performance by ignoring semi-hard and easy negative samples. Wehrmann *et al*. [25] proposed to use convolutional text encoders and perform convolutions over characters as opposed to words. They use an embedding matrix for characters and show significant reduction in number of paramters of the model.

You *et al*. [26] propose to use local context along with global loss to train the image embeddings. Their method represents each word in a caption by learning a word embedding matrix and perform series of 1-d convolutions over the individual words to get a final encoding of the caption.

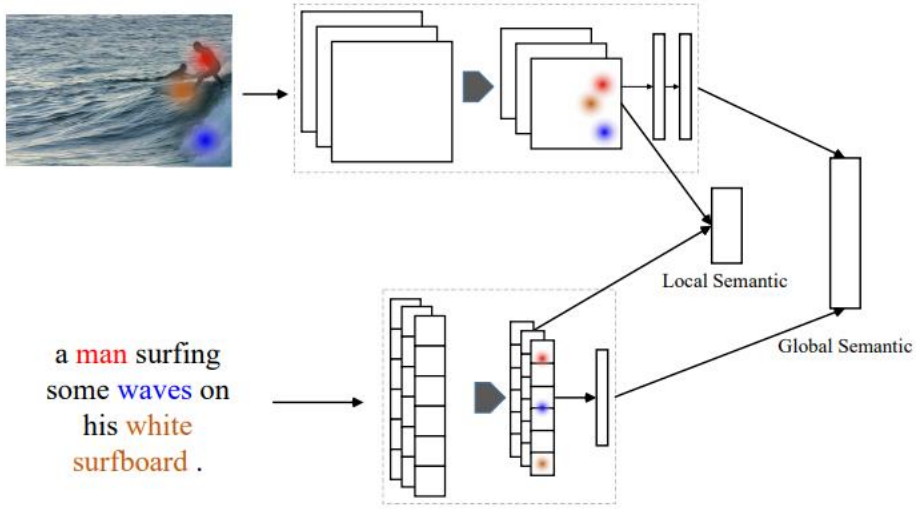


Figure 7 Convolutional Semantic Model [26].

Figure 7 shows their model which enforces a local loss between the intermediate convolutional layers of images and text. The margin based ranking loss is used for aligning both the local and global context.

Images consist of diverse content which can include objects as well as actions and attributes describing them. Most of the commonly occurring datasets like MSCOCO [7] and FLICKR 30K [27] contain only objects in the image and captions associated with them. Objects alone do not convey the semantic meaning of the image. Extending this idea, Yan *et al.* [28] built a vocabulary consisting of image categories, attributes and actions using the captions corresponding to each image. A caption describing the image contains more semantic information. Using the example "Two men are fighting on the road", semantic entities include "Two", "men" , "fighting" and "road". Nouns, adjectives and cardinal numbers are extracted from each caption and frequently occurring words are treated as discrete classes. Using these diverse classes, they train a multi-label CNN to identify the semantic concepts in the image.

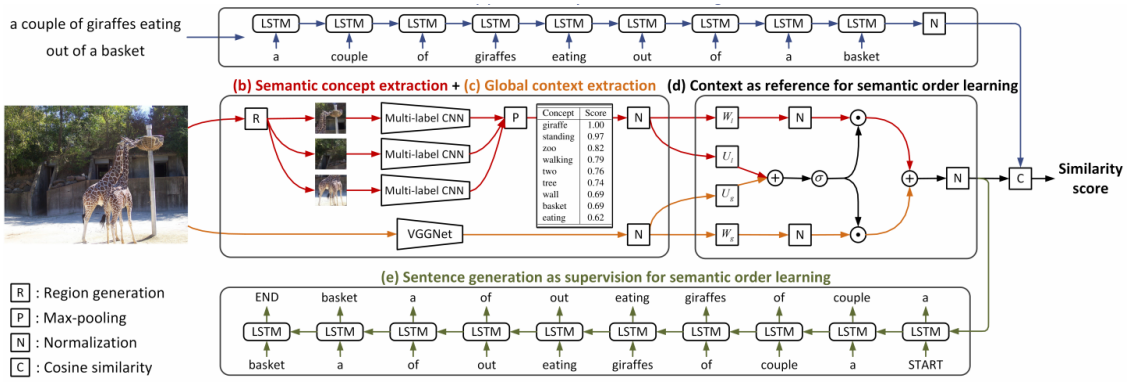


Figure 8 Image Sentence Matching using Multi-label CNN [28].

Figure 8 outlines their model where regions of an input image are passed to a multi-label CNN. The class probabilities of the multi-label CNN represent the distribution of semantic concepts in the image. A gated fusion unit is used which takes the semantic concepts and global features extracted by Resnet-152 [9] and outputs a fused vector representation which effectively weighs the importance of global and local features. The architecture of the gated fusion unit is similar to an LSTM cell where the sigmoid activation is applied to the linear combination of inputs which regulates the amount of information passed from input stage to the output stage. A Gated Recurrent Unit (GRU) is used as a sentence encoder where consecutive words in a caption are passed at each timestep of the network. A sentence generator is also used as supervision which ensures that the image can also generate the relevant caption. The sentence generation and the margin based ranking loss can effectively guide the image to better represent the content in the sentence during training. During inference, they extract '*r*' regions from each test image and pass each region to a multi-label CNN The value of '*r*' was set to 50. Output class probabilities vectors are obtained for all the regions and they are max-pooled. This results in a single vector which has the information of individual classes. The gated fusion unit combines the aggregate class probabilities vector which has the local context in the image along with the global feature vector extracted from Resnet-152 [9] to output the

final image embedding for the test image. This mechanism of learning semantic concepts and then matching it with the sentence embedding significantly improved the performance of retrieval.

Martin *et al*. [38] proposed to use selective pooling of the convolutional feature maps in the setting of a two branch network to enhance cross modal retrieval. Figure 9 shows their model where the selective pooling is applied at the pool block before the affine normalization of the image embedding.

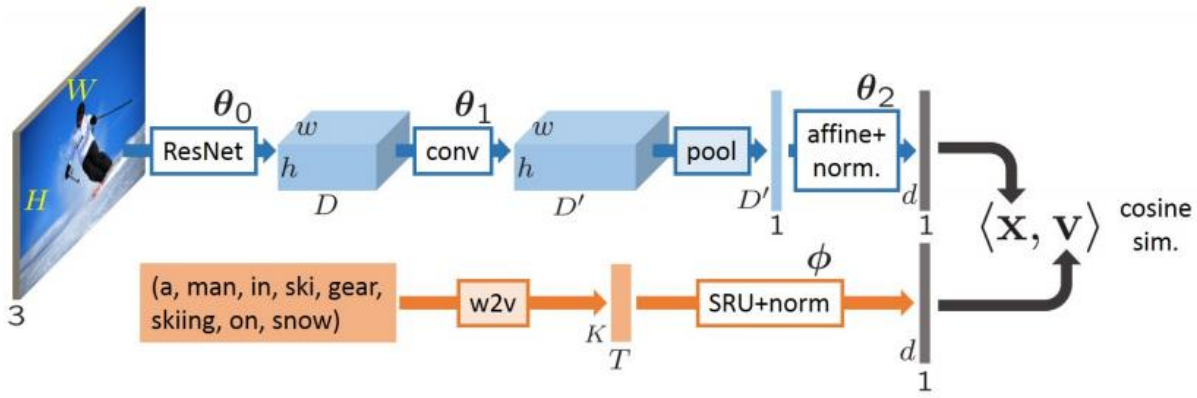


Figure 9 Selective pooling of convolutional feature maps for image-sentence matching [38].

Selective spatial pooling is given by (6):

$$h[k] = \max G(:,:,k) + \min G(:,:,k)\,,\ k = 1\ to\ D' \qquad (6)$$

where G is a convolutional feature map of size width x height x D'. D' represents the number of feature maps of last layer of Resnet-152 [9]. The selective spatial pooling can be considered as an aggregation of max pooling and min pooling. A simple recurrent unit (SRU) which is a 4-layer GRU network serves as a text encoder for captions. During training, all the

parameters of Resnet-152 [9], SRU and the embedding layers are learned by optimizing the margin based ranking loss between image and caption pairs.

Sah *et al*. [43] proposed a Common Vector Space (CVS) which brings similar concepts from different modalities closer in this space. They used different variants of metric learning loss functions [12, 15, 17] during training to achieve a common latent representation between images and text. One of the key difference between other methods is the way they infer the embeddings from CVS. During inference, they use the embeddings from either modality to reconstruct the original images using an image generator.

Chapter 3

Methodology

3.1 Baseline Model

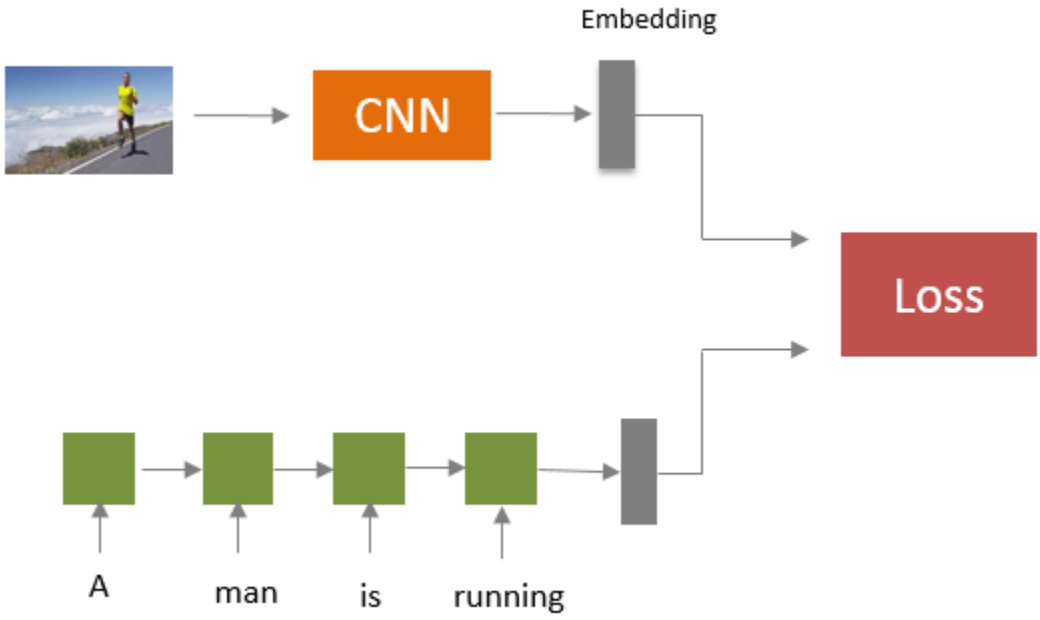


Figure 10 Baseline Model.

In order to establish a CVS between images and text, we need encoders which extract semantic information from individual modalities. Figure 10 shows the baseline architecture which is used for cross modal retrieval in this research. An input image is passed through a deep convolutional network [8, 9, 10] which extracts global features. These features are passed through a fully connected layer whose output is the vector representation of the image in the common embedding space. The sentence is encoded via GRU or LSTM and then passed into a fully connected layer. Margin based similarity loss is enforced on the image and text

embeddings which ensures similar concepts come closer and dissimilar concepts are pushed far apart by at least *margin* in the common embedding space.

Margin based Ranking Loss

Face recognition has seen significant progress in recent years and most of it can be attributed to metric learning loss functions that enhance the learning of the model. Several novel loss functions have been proposed [15, 17, 18, 19, 21], which exploit the batch to form exhaustive positive and negative pairs. Given a batch of samples, each sample is compared against all other samples in the batch. The number of triplets that can be formed in a batch is of the order $O(n^3)$. The number of contrastive pairs that can be formed in a batch is of the order $O(n^2)$. Optimizing over all these combinations is computationally infeasible and pose heavy memory constraints on fitting large models on standard GPUs. Sampling strategies such as hard and semi-hard negative mining have thus been proposed to mitigate this issue. Equation (7) shows an extension of triplet and lifted structured loss for cross modal tasks.

$$L_{\text{sim}} = \sum_m \sum_k \max\big(0, \alpha - S(i,c) + S(i,c_k)\big) + \sum_k \sum_m \max\big(0, \alpha - S(c,i) + S(c,i_k)\big) \quad (7)$$

where is the margin of separation of positive and negative pairs, *c* denotes a caption and *i* denotes an image. In (7), '*m'* denotes the total number of images and '*k'* denotes the total number of sentences in a batch. The first term in the equation is associated with caption retrieval where a single image is compared against all the '*k'* captions in the batch. The second term in the equation is associated with image retrieval where each caption in the batch is compared against all other '*m'* images in the batch. This loss essentially enforces the common embedding space to form distinct clusters for different entitites. The term *S(c, i)* computes the

similarity between a caption and an image and *S(i, c)* computes the similarity between an image and caption. This similarity can be a cosine similarity as shown in (8) or we can use the order violation penalty proposed in [23]. The order violation penalty enforces hierarchy of captions over image given by (9). It is always computed with the caption embedding being the anchor. The baseline model is a simplified model with all the necessary image and text pipelines.

$$S(i,c)_{cosine} = i.c^T \tag{8}$$

$$S(i,c)_{order} = max(0, |c| - |i|)^2 \tag{9}$$

where i, c denote the image and caption embeddings and $|i|$ denotes the absolute value of the image embedding.

3.2 Show, Translate and Tell (STT)

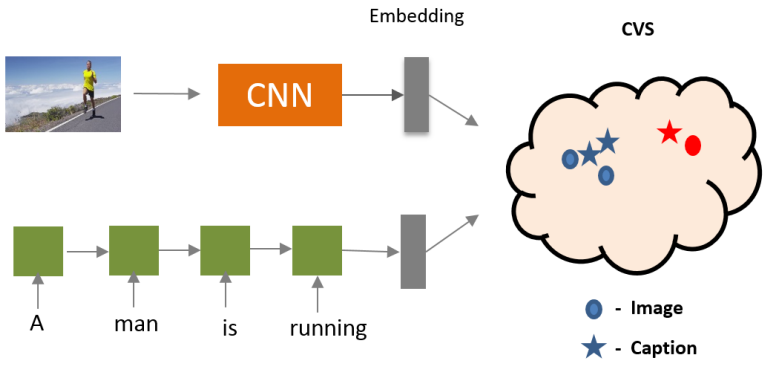


Figure 11 Common Vector Space (CVS) of Images and Text.

Figure 11 shows the CVS of images and text embeddings. CNNs and RNNs act as encoders for images and text. Images and captions which are semantically similar are mapped closer in this CVS. For example, in Figure 11, the image of a man running is eqiuvalently described by the sentence "A man is running". They are treated as positive pairs marked by the blue circle and star in the Figure 11. The margin based ranking loss in (7) brings these positive pairs closer and maps all the other negative pairs denoted by red circle and red star far apart by atleast *margin*.

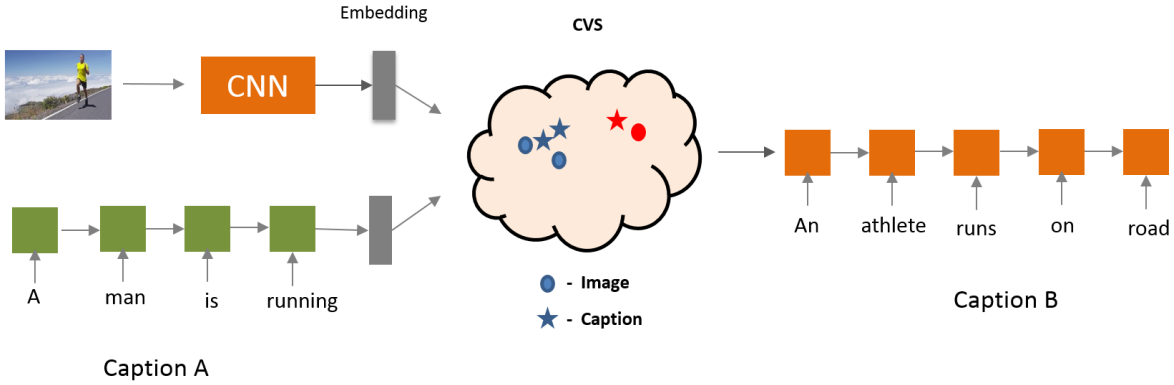


Figure 12 Show, Translate and Tell.

The baseline model constructs a CVS where the relationships between images and text are expressed in terms of the similarity score between their respective embeddings. CVS is a continuous space- without training, data points corresponding to images and text from the original dataset would be mapped arbitrarily. Training CVS forms dense clusters of matching images and captions. In order to explore the information contained in these CVS data points, we need to decode them into meaningful representations. Modality specific decoders can generate either images or captions. One of the early approaches related to this idea was proposed by Sah *et. al* [43] where they decode CVS embeddings using an image generator to generate images. Their method does not train the decoder along with the encoders during

training. They use a pre-trained image generator [44] as decoder which accepts a 4096 dimensional vector as input to generate images. This limits the interpretability of CVS in that the underlying distribution of CVS might be different from the input distribution that is expected by the image generator. More often, the generated images experienced the phenomenon of mode collapse. Inspired by [43] and image captioning models [29, 30, 31, 32], we propose Show, Translate and Tell (STT) which represents images and text in the CVS and also decodes the embeddings into captions by using an RNN. Figure 12 shows the schematic of the proposed model. STT offers a simple way to infer the embeddings in CVS by using an RNN as a decoder which is trained along with the image and text encoders. This ensures that decoder is aware of the distribution of the CVS embeddings. Since the output of the decoder was intended to be paraphrase captions, RNN was the preferred choice to generate these sentences.

During training, a single sample constitutes an image and two captions (caption A and caption B) as shown in the Figure 12. The captions describe the contents of an image. Features are extracted from image and caption 'A' using deep convolutional and recurrent networks. These features are projected into a CVS which aligns similar images and captions.

Caption 'B' which is always semantically similar to caption 'A' is used to enhance the overall quality of the model. The left side of the model in Figure 12 comprises of encoder models which encode a modality into its corresponding representation. The right side of the model is a recurrent neural network which acts as a decoder for both the image and caption A. Individual components of the model can be viewed in Figure 13.

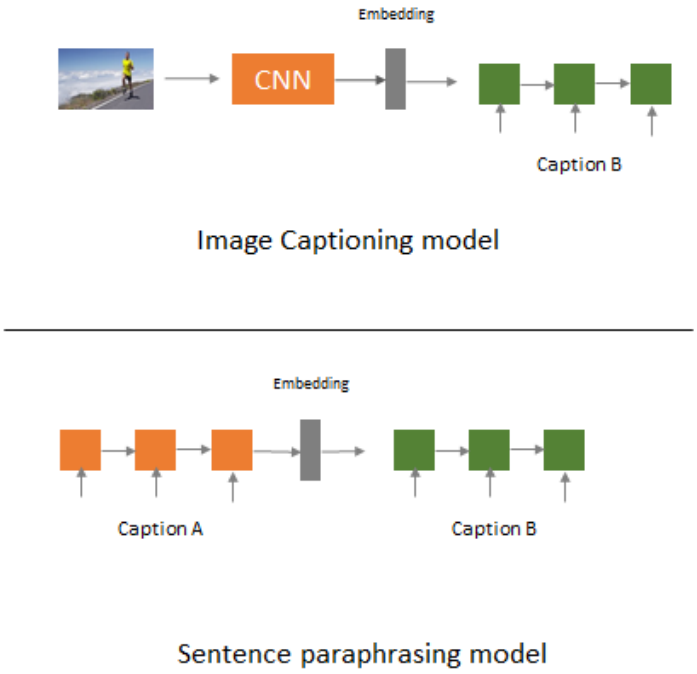


Figure 13 Image Captioner and Sentence Paraphraser.

Image sentence matching is closely related to sentence paraphrasing and image captioning. In image captioning models, an input image is projected into its feature space and passed to an RNN. During training, the RNN tries to correlate words in the sentence with the objects and actions in the image. A vocabulary of words is built using the most frequently occurring words in the captions. Each word in the vocabulary is encoded into a vector representation by a randomly initialized word embedding matrix. The word embedding matrix acts as a look-up table for the words in the caption. During training, the embedding matrix is also learned along with the weights of RNN and CNN. This ensures the word embedding matrix accurately learns the relationships between words and the underlying context within a sentence.

During training, the grountruth words are passed as input at current timestep instead of the word predicted by previous timestep. Cross-entropy loss between the predicted words from each timestep of RNN and the groundtruth sentence is used to optimize the parameters of the model. During testing, an image is passed through the network along with a start token for decoding the words in the sentence. Equation (10) denotes the loss that is used to train an image captioner.

$$L_{IC} = -\sum_{t=1}^{N} \log P(S_t|I;\ \theta) \tag{10}$$

where $P(S_t)$ is the probability of observing the correct word S_t at time t, denotes the paramters of the model and I denotes the image features.

Sentence paraphrasing models transforms a caption 'A' into caption 'B' which is semantically similar to caption 'A'. The words in the caption are encoded into vector representations by the embedding matrix. Sentence paraphrasing models are modeled in encoder-decoder framework where both encoder and decoder use recurrent neural networks. During testing, the input to the model is the encoded sentence representation by the encoder along with the start token. Equation (11) denotes the cross entropy loss used to train the sentence paraphraser.

$$L_{para} = -\sum_{t=1}^{N} \log P(S_t|E;\ \theta) \tag{11}$$

where $P_t(S_t)$ is the probability of observing the correct word S_t at time t, E denotes the encoder representation of the sentence and θ denotes the paramters of the model.

The sentence paraphrasing model ensures the two sentences are closer in the embedding space. It also ensures the encoded representation captures the semantic context which can be decoded

into a similar representation. Combining the benefits of the image captioning model and sentence paraphrasing model, Figure 12 is a unified model which can perform three different tasks namely image-caption retrieval, image captioning and sentence paraphrasing.
Equation (12) shows the loss for the unified model.

$$L = \lambda_1 L_{IC} + \lambda_2 L_{para} + \lambda_3 L_{sim} \quad (12)$$

where L_{IC}, L_{para} and L_{sim} correspond to the image captioning, sentence paraphrasing and similarity loss respectively. λ_1, λ_2 and λ_3 are the weights for each of the components of the above loss functions.

3.3 STT with Attention

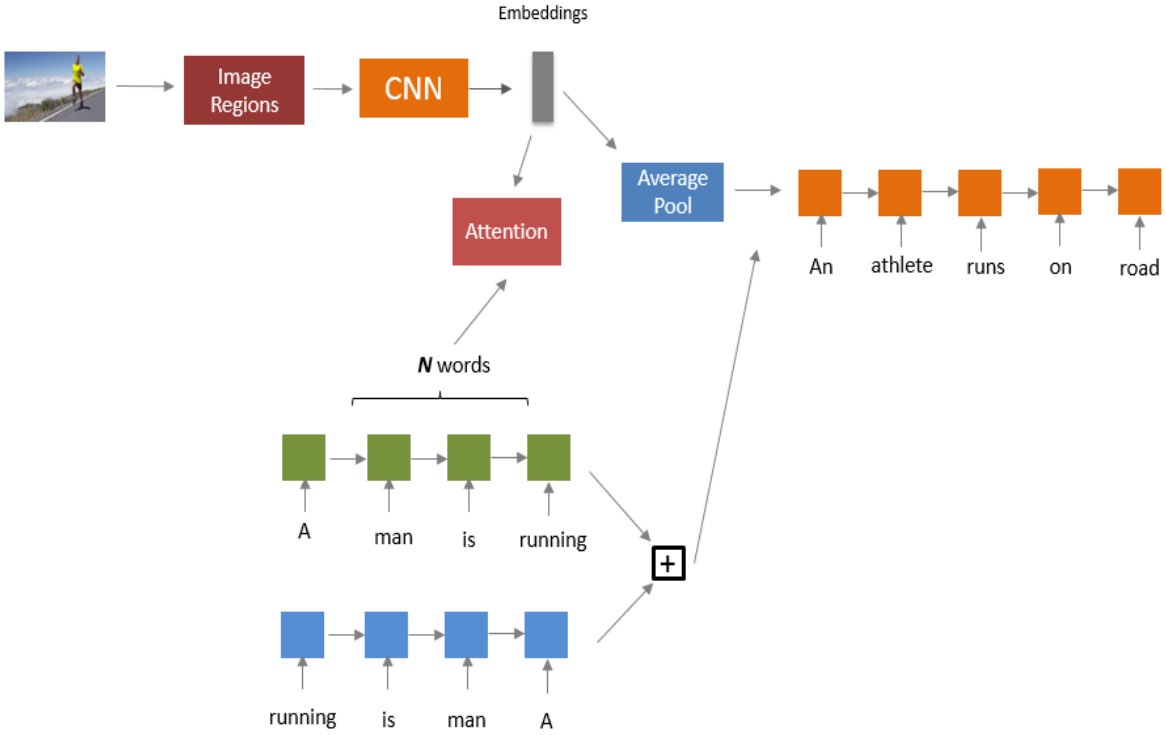


Figure 14 Show, Translate and Tell model with Attention.

Figure 14 describes the Show, Translate and Tell model with attention. The input image is passed through an object detector which outputs region proposals with objectness score in the image. The object detector used in this process is Faster R-CNN [36] which is a two-stage object detection network. In the first stage, it outputs region proposals with amount of objectness in each of them. These proposals are later refined in the second stage and bounding box regressor head localizes the objects in the image. The number of proposals after the first stage is 300. We consider a subset of '***M***' of these proposals based on their objectness score and extract the portions in the original image. These extracted proposals are passed through a pre-trained Resnet-152 layer CNN [9] and ***'M'*** regional embeddings are obtained. In

our experiments, we set the value of '***M***' to be 36. For more local information and semantic understanding of the contents in the image, we follow [37] to introduce attention between the region proposals in the image and individual words in the sentence. The similarity matrix introduced by [37] calculates the similarity between regions in the image and words in the sentence and is given by

$$s_{ij} = \frac{v_i^T e_j}{||v_i||\,||e_j||} \tag{13}$$

where s_{ij} is the similarity between i^{th} region (v_i) and j^{th} word (e_j). Based on the above similarity matrix, an attended sentence vector is calculated as

$$\mathrm{a}_\mathrm{i}^t = \sum_{j=0}^{n} \alpha_{ij} e_j \tag{14}$$

where

$$\alpha_{ij} = \frac{\exp(\lambda_1 s_{ij})}{\sum_{j=1}^{n} \exp(\lambda_1 s_{ij})} \tag{15}$$

α_{ij} is the attention weights which calculates the importance of each word in the sentence with respect to i^{th} region. The similarity between an image and text is then defined as a mean of similarity between image regions and attended sentence vectors. Equation 16 shows the aggregate similarity score between an image and a text.

$$S(I,T) = \frac{\sum_{i=1}^{k} R(v_i, a_i^t)}{k} \tag{16}$$

where $R(v_i, a_i^t)$ is the similarity of i^{th} image region and attended sentence vector given by

$$R(v_i,\ a_i^t) = \frac{v_i^T a_i^T}{|v_i|\,|a_i^t|} \tag{17}$$

This way of calculating similarity between regions in the image and words can be directly plugged into (7) where the similarity of the negative pairs is reduced thereby bringing the positive samples closer in the common embedding space. One difference between this attention model and STT is that a single image is represented by multiple region embeddings rather than a single feature vector. We add an average pooling layer which aggregates these multiple region embeddings into a single representation which can be used as an input to the decoder. The joint training of the decoder and individual image and text encoders along with the attention model helps in aligning the image regions with the individual words and also in generating high quality sentences.

Chapter 4

Implementation and Results

4.1 Datasets

Some of the popular cross modal datasets which include images and captions associated with them include

- MSR-VTT [33]
- Caltech-UCSD Birds 200 [4]
- Flowers 102 dataset [3]
- MSCOCO [7]
- FLICKR 30K [27]

MSR-VTT [33] is a large scale video to text dataset which bridges video and language. It has comprehensive categories and diverse video content which can be used for video retrieval, event detection tasks.

Caltech-UCSD Birds 200 [4] is a medium scale dataset consisting of 200 categories of birds along with attributes for each image. Each image is also annotated with 10 captions which describe the content in the image.

Flowers 102 [3] dataset is a dataset by University of Oxford which consists of 102 different categories of flowers with 10 captions associated with each image.

MSCOCO [7] dataset is a large-scale dataset comprised of common objects that are found in nature. It is widely used for multi-label classification, object detection, semantic segmentation and cross modal retrieval.

Table 1 Summary of cross-modal datasets.

Statistics	MSR-VTT	Caltech Birds	Flowers 102	MSCOCO	FLICKR 30K
Data	Video and Text	Images, Captions	Images, Captions	Images, Captions	Images, Captions
# train samples	260K	8855	7034	82,783	29,783
# validation samples	_	2933	1155	40,504	1000

This work focuses on bridging natural image content and corresponding descriptions. Caltech-UCSD Birds dataset [4] and Flowers 102 dataset [3] are very specific to birds and flowers and the diversity of the dataset in terms of datasets and semantic content is limited. These datasets are more oriented towards zero shot image retrieval and classification. MSR-VTT [33] is more suitable for temporal video analysis, video segmentation and captioning. In this work, MSCOCO [7] and FLICKR 30K [27] datasets are explored which contain real-world images with short descriptions associated with them.

4.2 Training Details

Each image is passed through a Resnet-152 CNN [9] and features are extracted from the global average pooling layer. For the embedding network, we use a single fully connected layer. Training is performed in multiple stages. In the first stage, we pre-compute the features from Resnet-152 [9] and train the image embedding and sentence encoder from scratch for 15 epochs with a learning rate of 0.0002 and lower the learning rate to 0.00002 for the next 15

epochs. We use Adam optimizer for optimizing the parameters of the model. Once we have the model trained with the precomputed features, we finetune the Resnet-152 CNN along with embedding layers and sentence encoder with a learning rate 0.00002 for another 15 epochs. We found the model to be highly sensitive to the learning rate and higher learning rates often led to model getting stuck at local minimum. We use Tensorflow deep learning framework for all our experiments.

For the sentence representation, we use a 1-layer GRU network. The hidden dimension of the GRU was set to 1024. We experimented by stacking more layers, but there was no significant improvement by introducing more parameters. This complemented with our usage of only one fully connected layer for generating embeddings. The vocabulary of words was built by counting the frequency of all the words in the captions present in the dataset. A word is considered to exist in vocabulary if the frequency of its occurrence is greater than three. The size of the vocabulary is 26,375 words. The word embedding dimension was set to 300.

For the margin based ranking loss, we set the margin to 0.05 and we use order violation penalty [23] for computing the similarity metric. The batch size is set to 128, so the number of contrastive examples for each matching pair would be 127. We employ hard negative mining where we only consider the hardest negative distance instead of aggregating all the negative distances. We noticed a significant performance improvement in retrieval with hard negative mining strategy.

Table 2 shows the statistics for MSCOCO data that were used in our experiments.

Table 2 MSCOCO statistics.

MSCOCO	Baseline Model	STT Model
Training images	113,287 (train+val)	113,287
Number of captions	565,435	565,435
Total number of samples	565,435	11,308,700
Test images	5000	5000

For the baseline model, a sample constitutes an image and caption. Since each image is associated with a set of five captions, the training set of images constitute 565,435 samples in total. For the STT model, a sample constitutes an image and two similar captions. Since we have five captions associated with each image, there can be 20 different combinations. Hence, the total number of samples for the STT model would be 11,308,700.

Table 3 shows the statistics for FLICKR 30K dataset that we used in our experiments.

Table 3 FLICKR 30K statistics.

FLICKR 30K	Baseline Model	STT Model
Training images	29,783	29,783
Number of captions	148,915	148,915
Total number of samples	148,915	2,978,300
Test images	1000	1000

For the baseline model, a sample constitutes an image and caption. Since each image is associated with a set of five captions, the training set of images constitute 148,915 samples in total. For the STT model, a sample constitutes an image and two similar captions. Since we

have five captions associated with each image, there can be 20 different combinations. Hence, the total number of samples for the STT model would be 2,978,300.

4.3 Evaluation Metrics

The following evaluation metrics are used as a standard to compare the performance of cross-modal retrieval.

4.3.1 Recall@*K*

Recall@*K* computes the recall at different values of *K*. It is a metric which computes if the rank of retrieved sentence is within the top *K* ranks. All the test set images and their associated captions are passed through the model and their embeddings are extracted. Each image embedding in the test set is compared with all other caption embeddings and similarity metric is computed. The similarity is sorted in descending order and appropriately all other captions are ranked. If the rank of the groundtruth sentence is within the top *K* ranks, the caption is counted as a positive retrieval. Typical values of *K* are 1, 5 and 10. The overall recall score is the percentage of samples within the top *K* ranks.

For the image retrieval, we rank all images in the test set with every caption. The images are sorted in descending order with respect to the similarity with respect to query caption and ranked. If the rank of the groundtruth image is within the top *K* ranks, the image is counted as a positive retrieval.

4.3.2 Median rank

After computing the ranks for each samples, the median of these ranks is computed. Ideally the median value should be 1 which implies all the samples should be correctly retrieved.

4.3.3 Mean rank

Mean rank is the mean of the ranks of all samples in the test set. The mean rank should also be equal to 1 in the perfect scenario where all the samples are correctly retrieved.

4.4 Baseline Results

4.4.1 MSCOCO

Table 4 Results of MSCOCO sentence retrieval using baseline model.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
Baseline	1 FC	1024	55.5	85.2	92.3	1	4.6
Baseline	1 FC	2048	56.4	85.2	92.6	1	4.7

Table 5 Results of Image retrieval on MSCOCO test set using baseline model.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
Baseline	1 FC	1024	41.4	75.1	85.9	2	12.2
Baseline	1 FC	2048	42.5	76	86.5	2	11.6

Tables 4 and 5 show the results of sentence retrieval and image retrieval on MSCOCO dataset using the baseline model. The model configuration indicated in the tables is '1 FC' which indicates that the image and text branches consist of one fully connected layer each. The column '*Emb dim*' indicates the size of the image and text embeddings. The network architecture comprises of a Resnet-152 layer CNN for encoding images and GRU recurrent

network for encoding captions. From Table 4 and Table 5, we can conclude that increasing the embedding dimension does not significantly affect the retrieval performance. There is a slight improvement in other metrics.

The recall scores increase as we increase the values of *K.* This implies that there are significant number of retrievals within the top 10 ranks. The recall values of sentence retrieval are comparatively higher than image retrieval due to the fact that each image is associated with five different captions. A sentence retrieval is considered a positive retrieval if the retrieved sentence belongs to any of the five associated captions for the image. In the image retrieval case, each caption is associated with only one image which makes the problem more challenging. This is evident from the R@1 scores of sentence and image retrieval which are 55.5 and 41.4 respectively.

4.4.2 FLICKR 30K

Table 6 Results of Sentence Retrieval using Baseline model on FLICKR 30K dataset.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
Baseline	1 FC	1024	40.2	67.1	79.4	2	15.442
Baseline	1 FC	2048	38.4	67.4	77.5	2	13.4

Table 6 shows the results of sentence retrieval using baseline model on FLICKR 30K dataset. The recall scores are comparatively lower than that of MSCOCO due to fewer samples in the dataset. The model quickly overfits the training data hurting the performance on the test set. We tackle overfitting by monitoring the model's performance on validation data and choosing the best model accordingly.

Table 7 Results of Image Retrieval using Baseline model on FLICKR 30K dataset.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
Baseline	1 FC	1024	27.42	55.58	67.9	4	28.98
Baseline	1 FC	2048	27.1	55.9	68.2	4	24.5

Table 7 shows the results of image retrieval using Baseline model on FLICKR 30K dataset. The recall scores are considerably lower compared to MSCOCO due to less number of samples. From Tables 6 and 7, it is clear that increasing the size of embedding does not help the performance of the retrieval model.

4.5 Results of STT Model

4.5.1 MSCOCO

Table 8 STT results on MSCOCO for Sentence Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT	1 FC	1024	54.7	83.6	92.1	1	4.5
STT	1 FC	2048	55.1	83.5	91.8	1	4.5

Table 8 shows the results of STT on MSCOCO for sentence retrieval. The model configuration indicated in the tables is '1 FC' which indicates that the image and text branches consist of one fully connected layer each. The column '*Emb dim*' indicates the size of the image and text embeddings. The recall scores seem improve by 0.4% when the embedding dimension is increased. The scores are low compared to the baseline results. One reason might

be that the model is overfitting the data since the new dataset for STT is 20× the original image-caption pairs with many repetitive pairs as indicated in Table 3.

Table 9 STT results on MSCOCO for Image Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT	1 FC	1024	41	74.8	86	2	9
STT	1 FC	2048	41.3	75.2	86	2	9.3

Table 9 shows the results of STT on MSCOCO for image retrieval. The recall scores do not seem to improve significantly with the increase in embedding dimension.

Image Captioning

The STT model is flexible and can perform diverse tasks. The top part of STT shown in Figure 13 can effectively be used as an image captioner. The task of image-sentence matching is performed by representing images and text close to each other in the common embedding space. This high dimensional space is comprised of many naturally occurring images and text that lie outside of the dataset. Our image captioner effectively generates sentences which lie near the vicinity of the corresponding images. Table 10 shows the results of image captioning on MSCOCO 1K test set.

Table 10 Image Captioning Results of STT model on MSCOCO 1k test set.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT	1024	0.683	0.506	0.362	0.259	0.236	0.850
STT	2048	0.671	0.492	0.351	0.250	0.232	0.822

Table 10 indicates that the STT model is able to achieve good image captioning scores. The effect of embedding dimension is clearly less significant in image captioning when compared to cross modal retrieval.

Sentence Paraphrasing

The task of sentence paraphrasing model involves generating a paraphrase which is semantically similar to the input sentence. This task is particularly challenging due to the fact that a sentence can be described in many ways. The generated sentence should not only capture the context of a sentence but it should also be syntactically different from the input sentence. We evaluate our STT model on the task of sentence paraphrasing. Table 11 shows the result of sentence paraphrasing on MSCOCO 1K test set.

Table 11 Sentence paraphrasing results on MSCOCO 1K test set using STT model.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT	1024	0.744	0.578	0.435	0.324	0.275	1.10
STT	2048	0.734	0.568	0.426	0.317	0.270	1.069

From Table 11, it is clear that the STT model can generalize well on sentence paraphrasing tasks. It is able to obtain good scores which can be attributed to the fact that we are jointly training the model on sentence paraphrases which contain more context. The sentence decoder in STT model effectively makes sure the embeddings in the common vector space have semantic meaning and enables captioning and paraphrasing applications.

Visualizations

1) Good examples

Captioning : a group of people riding bikes down a street

Paraphrasing : a group of people riding bikes down a street
a man riding a bike down a street next to a traffic light

Retrieved captions

- bike riders passing Burger King in city street
- A group of bicyclists are riding in the bike lane .
- Bicyclists on a city street , most not using the bike lane

Groundtruth captions

- people on bicycles ride down a busy street
- A group of people are riding bikes down the street in a bike lane
- bike riders passing Burger King in city street
- A group of bicyclists are riding in the bike lane .
- Bicyclists on a city street , most not using the bike lane

Figure 15 Sample STT output on MSCOCO.

Captioning : a man is standing on a dirt road

Paraphrasing : a person riding a motorcycle on a dirt road
a man riding a motorcycle down a dirt road

Retrieved captions

- A man with a red helmet on a small moped on a dirt road .
- A man in a red shirt and a red hat is on a motorcycle on a hill side .
- Man riding a motor bike on a dirt road on the countryside

Groundtruth captions

- A man with a red helmet on a small moped on a dirt road.
- Man riding a motor bike on a dirt road on the countryside.
- A man riding on the back of a motorcycle.
- A dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains.
- A man in a red shirt and a red hat is on a motorcycle on a hill side.

Figure 16 Sample STT output on MSCOCO.

Figures 15 and 16 show STT outputs on a sample images. From the figures, we can observe that the top 3 retrieved sentences are a part of the groundtruth sentences for the image. This indicates that the STT model was able to retrieve sentences very well. The image captioning and sentence paraphrasing results also describe the image well.

2) Bad examples

Captioning : a wooden table topped with lots of wooden benches

Paraphrasing : a wooden table topped with lots of pairs of scissors
a wooden table with a wooden cutting board and knives

Retrieved captions

- A group of animals walking in the grass next to a road
- A giraffe and zebras mingle as cars drive out of an animal park .
- a giraffe standing in the middle of some zebras

Groundtruth captions

- Multiple wooden spoons are shown on a table top .
- A table surrounded by chairs and filled with cooking utensils .
- Wooden spoons laid out across a kitchen table .
- Wooden spoons and forks are all over a table .
- A table and chairs with wooden kitchen tools on top .

Figure 17 Sample STT output on MSCOCO.

Figure 17 shows an example where STT model failed to retrieve the right captions. The retrieved captions describe content related to a group of animals and giraffes. This might be due to the texture formed by the wooden spoons on the table as well as resulting color similar to the color of giraffes. The captioning and paraphrasing show good results even though the retrieval failed.

Captioning : a wooden table topped with lots of wooden benches

Paraphrasing : a wooden table topped with lots of pairs of scissors
a wooden table with a wooden cutting board and knives

Retrieved captions

- The small bathroom has wooden cabinets around the sink .
- A large white bathroom with white cabinets and double sinks
- A bathroom with a sink , toilet , and vanity .

Groundtruth captions

- The bathroom is clean and ready to be used .
- a small bathroom with a sink and a toilet .
- A toilet , sink and mirror in the bathroom
- Doorway view into bathroom with a sink and toilet .
- A white toilet sitting next to a sink .

Figure 18 Sample STT output on MSCOCO.

Figure 18 shows an example of ambiguous retrieval. Although the retrieved sentences for the query image are semantically related to the image, they do not belong to the groundtruth sentences which makes this a negative retrieval. The retrieved captions might be related to another image with similar content. This example depicts that the cross-modal retrieval is very challenging when there is high overlap of semantic content in the dataset.

4.5.2 FLICKR 30K

Table 12 Sentence Retrieval results on FLICKR 30K dataset using STT model.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT	1 FC	1024	38.9	66.9	78.4	3	13.3
STT	1 FC	2048	38.4	67.4	77.5	2	13.4

Table 13 Results of Image Retrieval on FLICKR 30K dataset using STT model.

Variant	Model	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT	1 FC	1024	27.2	55.4	68.4	4	27.6
STT	1 FC	2048	27.1	55.9	68.2	4	24.5

Table 12 and 13 show the results of sentence retrieval and image retrieval on FLICKR 30K using STT model. STT model's performance is lower than the baseline model for retrieval but still shows strong performance. STT results on FLICKR 30K [27] are consistent with MSCOCO [7].

Table 14 Results of Image Captioning on FLICKR 30K using STT model.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT	1024	0.513	0.330	0.204	0.129	0.178	0.252
STT	2048	0.508	0.323	0.198	0.124	0.167	0.216

Table 15 Results of Sentence Paraphrasing on FLICKR 30K using STT model.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT	1024	0.569	0.394	0.262	0.176	0.217	0.398
STT	2048	0.548	0.364	0.233	0.151	0.189	0.292

Tables 14 and 15 show the results of image captioning and sentence paraphrasing on FLICKR 30K [27] using STT model. As the embedding dimension is increased, the scores decreased. This indicates that increasing the embedding size is not always beneficial.

Visualizations

1) Good examples

Captioning : a man in a black wetsuit is surfing on a surfboard a small wave

Paraphrasing : a surfer is jumping off a wave a wave the surfer
a man in a wetsuit is surfing a wave a man in a black shirt him

Retrieved captions

- A man surfing in the ocean .
- A man surfing in the ocean
- One male surfer doing a turn on a surfboard in the ocean .

Groundtruth captions

- A surfer balances on a surfboard while another surfer is not on his board in the background .
- A man in a wetsuit trying to balance on a surfboard .
- Two men are in the water using waterskiing equipment
- A man tries to keep balance while surfing
- A man surfing in the ocean

Figure 19 Sample STT output on FLICKR 30K.

Figure 19 shows an example of a good retrieval for a query image. The sentence '*A man surfing in the ocean*' is repeated twice in the dataset and they belong to two different image samples. Captioning results describe the image in more detail although the syntax is slightly affected. The paraphrasing also outputs good results and shows good diversity.

2) Bad examples

Captioning : a man and a woman are sitting at a table with a red tablecloth and a red tablecloth

Paraphrasing : a woman and a young girl are playing monopoly on a table a table a
a group of people are playing monopoly on a table

Retrieved captions

- A family sits down at a table to play a board game together .
- A group of adults plays a board game at a table
- A man and a woman are making two bowls of salad together .

Groundtruth captions

- Two women and two out of frame people playing a game of Monopoly .
- The girl in red is taking her turn in the Monopoly game .
- Three individuals are playing a game of Monopoly .
- A group of people are playing Monopoly .
- Three people are playing monopoly .

Figure 20 Sample STT output on FLICKR 30K dataset.

Figure 20 shows the results of a failed retrieval on FLICKR 30K [27] dataset. The retrieved captions do not belong to set of groundtruth captions. However, the retrieved captions describe the image accurately. This ambiguity can be attributed to the positive and negative pairing in the dataset during training. Since there is a significant overlap between some samples in the training data, the strict definition of each sample being a negative to all other samples in the dataset results in such ambigious scenarios.

4.5.3 Cross Domain Evaluation of STT model

In order to explore the generalization performance of STT model on other datasets, we perform cross-domain evaluation. We evaluate the STT model trained on MSCOCO [7], on FLICKR 30K [27] and vice-versa.

Table 16 Transfer learning results of STT model on Sentence Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10
STT	MSCOCO-FLICKR 30K	1024	32.9	57.4	67.4
STT	FLICKR 30K-MSCOCO	1024	24.8	50.5	62.4

Table 17 Transfer learning results of STT model on Image Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10
STT	MSCOCO-FLICKR 30K	1024	21.1	43.5	55.4
STT	FLICKR 30K-MSCOCO	1024	17	41.3	55.3

Tables 16 and 17 show transfer learning results of STT model on sentence and image retrieval. The model configuration ‘MSCOCO-FLICKR 30K’ indicates that the model was trained on MSCOCO [7] and evaluated on FLICKR 30K [27] dataset. The model ‘MSCOCO-FLICKR 30K’ performs well on FLICKR 30K [27] dataset compared to ‘FLICKR 30K-MSCOCO’ on

MSCOCO [7] dataset. This can be due to the fact that the MSCOCO [7] dataset is a much larger dataset as compared to FLICKR 30K [27] dataset. Since MSCOCO [7] and FLICKR 30K [27] have similar type of objects and content in the images, transfer learning is a good mechanism to evaluate the overall performance of the model.

4.6 Results of STT Model with Attention

4.6.1 MSCOCO

Table 18 Results of Sentence Retrieval on MSCOCO dataset using STT with Attention.

Variant	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT-ATT	1024	64.9	91	96.8	1	2.5

Table 19 Results of Image Retrieval on MSCOCO dataset using STT with Attention.

Variant	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT-ATT	1024	49.8	83	91.6	1	5.6

Tables 18 and 19 show the results of sentence and image retrieval on MSCOCO using STT model with attention (indicated by STT-ATT in the tables). As observed by [37], the retrieval scores show a significant improvement with attention. This concludes that cross-modal retrieval is a challenging task which requires fine-grained matching between images and captions.

Table 20 shows the results of image captioning on MSCOCO using STT model with attention. The table clearly shows improvement in the captioning scores over the STT model. The main difference between STT and STT with attention is the use of region proposals which

have local information about the objects in the image. The improvement in B@1 score with respect to STT model is 2.3%.

Table 20 Image captioning results on MSCOCO 1K test set using STT with attention.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT-ATT	1024	0.706	0.530	0.385	0.279	0.246	0.908

Table 21 Sentence paraphrasing results on MSCOCO 1K test set using STT with Attention.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT-ATT	1024	0.747	0.581	0.436	0.326	0.272	1.098

Table 21 shows the results of sentence paraphrasing on MSCOCO 1K test set using STT model with attention. The results are also complementary to image captioning results and show improvement over the STT.

Visualizations

1) Good examples

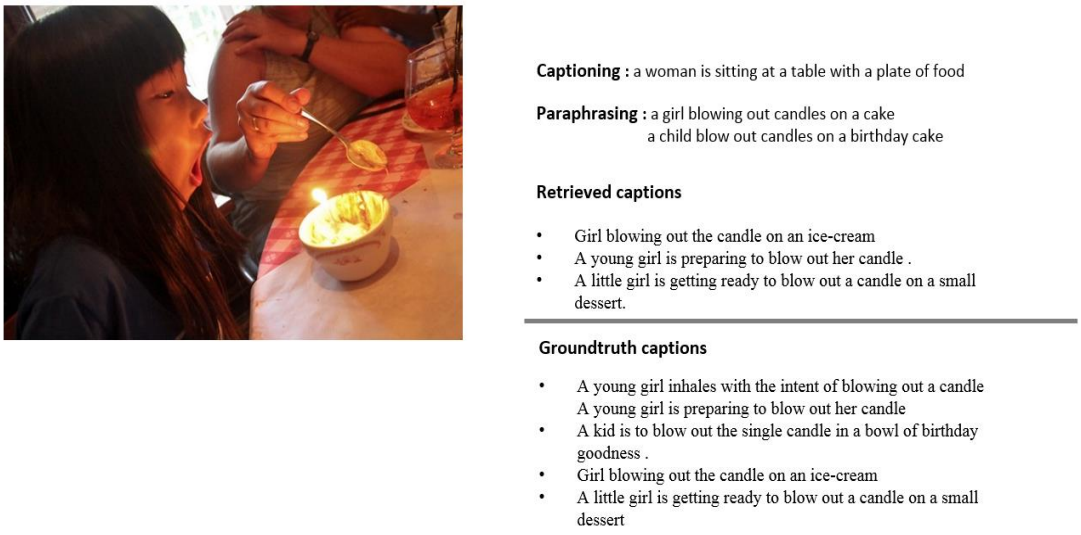
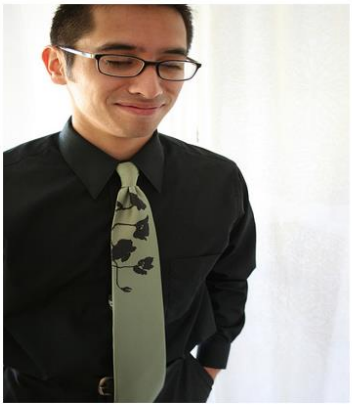


Figure 21 Sample output of STT-ATT model on MSCOCO.



Captioning : a man wearing a suit and tie standing in a room

Paraphrasing : a man wearing a green shirt and a tie
a man wearing a tie and a shirt smiling

Retrieved captions

- A man with glasses and his eyes closed dressed in a black shirt and a necktie
- A man wearing a suit and tie staring.
- A man with a tie and a suit .

Groundtruth captions

- A young man wearing black attire and a flowered tie is standing and smiling
- A man with glasses and his eyes closed dressed in a black shirt and a necktie
- A man in a green tie with his eyes closed
- Smiling man wearing black shirt and pale green tie
- A person that is dressed up very nicely .

Figure 22 Sample output of STT-ATT model on MSCOCO.

Figures 21 and 22 show good outputs of STT model with Attention on MSCOCO [7]. The retrieval results seem perfect and the outputs of captioning and paraphrasing captured the semantic content in the image.

2) Bad examples

Captioning : a bus parked in front of a tall building

Paraphrasing : two trucks are parked in a lot near a fire truck
people walking on a city street with buildings in the background

Retrieved captions

- People in a inner city courtyard watching a performance
- Train coming in to a station at the edge of a large city
- People standing in a long line at a train station

Groundtruth captions

- Two trucks that are sitting in the street .
- A blue delivery truck driving down a street .
- People walking past buildings and trucks on a cloudy day .
- a bench near a tree near a light pole
- A couple of delivery trucks parked next to a large building .

Figure 23 Sample STT-ATT output on MSCOCO.

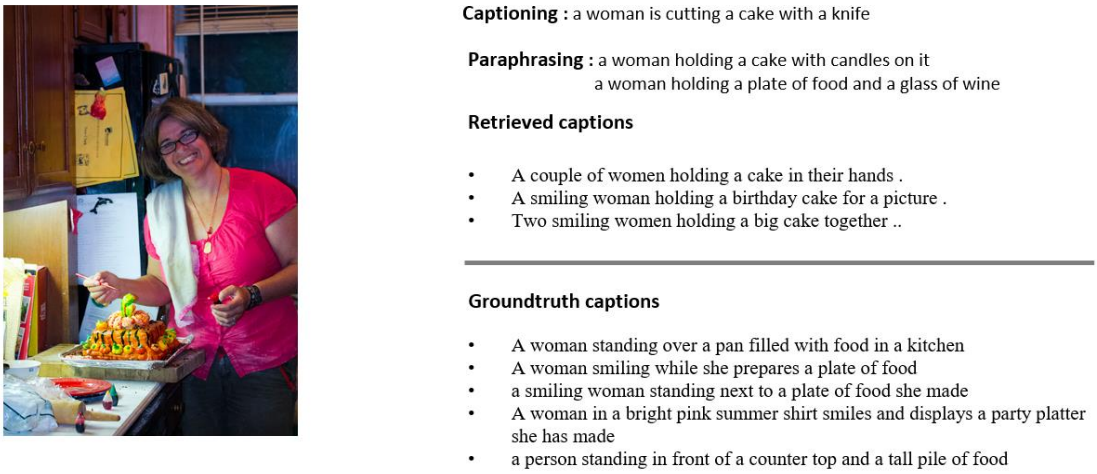


Figure 24 Sample STT-ATT output on MSCOCO.

Figures 23 and 24 show some failed retrievals of STT with Attention. In Figure 23, the model retrieves captions related to train and station. The image embeddings might not have been rich enough and the model confused the buildings with windows on a train. In Figure 24, the model is confused with the number of women in the picture. Although the retrieved captions reasonably describe the action of the woman, the groundtruth captions are different. These kind of examples are particularly challenging due to the high overlap between content in the data samples.

4.6.2 FLICKR 30K

Table 22 Results of Sentence Retrieval on FLICKR 30K dataset using STT with Attention.

Variant	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT-ATT	1024	59.2	83.5	91	1	6.6

Table 23 Results of Image Retrieval on FLICKR 30K dataset using STT with Attention.

Variant	Emb dim	R@1	R@5	R@10	Med R	Mean R
STT-ATT	1024	40.7	69.7	79	2	18.3

Tables 22 and 23 show the results of sentence and image retrieval on FLICKR 30K dataset using STT with Attention. The results show significant improvement compared to the STT model. This also shows the importance of attention in aligning modalities for datasets with fewer number of samples.

Table 24 Results of Image Captioning on FLICKR 30K using STT with Attention.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT-ATT	1024	0.611	0.427	0.293	0.203	0.193	0.442

Table 24 shows the results of image captioning on FLICKR 30K dataset using STT with attention. The scores improved as compared to the STT model without attention.

Table 25 shows the results of sentence paraphrasing on FLICKR 30K dataset using STT with Attention. The results also improve by adding the attention mechanism.

Table 25 Results of Sentence paraphrasing on FLICKR 30K using STT with Attention.

Variant	Emb dim	B@1	B@2	B@3	B@4	METEOR	CIDEr
STT-ATT	1024	0.673	0.493	0.353	0.252	0.221	0.573

Visualizations

1) Good examples

Captioning : dog runs through a lush lawn ahead

Paraphrasing : black and white dog runs in grass
brown and white dog runs across grass

Retrieved captions

- A dog runs on the green grass near a wooden fence.
- A Boston Terrier is running on lush green grass in front of a white fence.
- A brown dog with white paws is trotting through a field of green grass.

Groundtruth captions

- A black and white dog is running in a grassy garden surrounded by a white fence.
- A Boston Terrier is running on lush green grass in front of a white fence.
- A black and white dog is running through the grass.
- A dog runs on the green grass near a wooden fence.
- A Boston terrier is running in the grass.

Figure 25 Sample output of STT model with Attention on FLICKR 30K dataset.

Captioning : group of people in a snow race with trees and a man in background is

Paraphrasing : people pose for a picture in winter gear snow over a snowy mountain in
ride on motorized snow in the race in the city

Retrieved captions

- Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background.
- Five people wearing winter clothing, helmets, and ski goggles stand outside in the snow.
- Four people playing hockey in an ice rink.

Groundtruth captions

- Five snowmobile riders all wearing helmets and goggles line up in a snowy clearing in a forest in front of their snowmobiles; they are all wearing black snow pants and from left to right they are wearing a black coat, white coat, red coat, blue coat, and black coat.
- Five people wearing winter jackets and helmets stand in the snow, with snowmobiles in the background.
- Five people wearing winter clothing, helmets, and ski goggles stand outside in the snow.
- A group of snowmobile riders gather in the snow.
- Group gathered to go snowmobiling.

Figure 26 Sample output of STT model with Attention on FLICKR 30K dataset.

Figures 25 and 26 show the sample outputs of STT model with Attention on FLICKR 30K dataset. The retrieval results are perfect for these samples. In Figure 26, the results of captioning and paraphrasing are not accurate as they exhibit syntactic errors. This is due to the fact that FLICKR 30K dataset has fewer samples which makes it challenging for the model to learn the semantics and syntax of sentences.

2) Bad examples

Captioning : man with a mallet is carving a creature out of stone

Paraphrasing : woman playing the keyboard and singing on the street
woman playing a string instrument at an event by a crowd of people
on the stage

Retrieved captions

- An old woman working at a loom making cloth.
- An Asian woman wearing a Asian dress sitting among a group of cloths, with a woven basket on her lap.
- Indian man whittles in his own home.

Groundtruth captions

- A female harp player peers through the middle of her instrument while performing.
- A lady with dark hair is playing a harp.
- A pretty woman plays a harpsichord.
- A smiling woman playing the harp.
- A woman playing a harp.

Figure 27 Sample output of STT model with Attention on FLICKR 30K.

Captioning : child is splashing water at night under a bucket under a water

Paraphrasing : boy swinging on a tree from a tree stump by trees in the
group of people camping in a field outside on a sunny day playing
with their

Retrieved captions

- A girl drinking from a water fountain.
- A girl is drinking from a drinking fountain.
- A blond man is drinking from a public fountain.

Groundtruth captions

- A brown-haired child in green shoes swings on a swing in a park near the woods.
- The sun breaks through the trees as a child rides a swing.
- A child wearing Crocs sits on a swing in a wooded area.
- A kid swings with his feet up in the air in a forest.
- A child swings and the sunshine shines down on her.

Figure 28 Sample output of STT model with Attention on FLICKR 30K dataset.

Figures 27 and 28 show some failure cases of STT model with Attention on the FLICKR 30K dataset. In Figure 27, only the paraphrasing results capture the right information in the image. In Figure 28, the model failed on all three tasks, captioning, retrieval and paraphrasing. One possible reason might be the complexity of the image. These images have complicated content and the features might not be strong enough to produce good representations.

4.6.3 Cross Domain Evaluation of STT model with Attention

In order to explore the generalization performance of STT model with Attention on other datasets, we perform cross-domain evaluation. We evaluate the STT-ATT model trained on MSCOCO [7], on FLICKR 30K [27] and vice-versa.

Table 26 Transfer learning results of STT model with Attention on Sentence Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10
STT-ATT	MSCOCO-FLICKR 30K	1024	43.2	73.2	82.8
STT-ATT	FLICKR 30K-MSCOCO	1024	31.7	58.2	70.5

Table 27 Transfer learning results of STT model with Attention on Image Retrieval.

Variant	Model	Emb dim	R@1	R@5	R@10
STT-ATT	MSCOCO-FLICKR 30K	1024	35.1	61.3	72
STT-ATT	FLICKR 30K-MSCOCO	1024	22.5	50.2	64.1

Tables 26 and 27 show the transfer learning results of STT model with Attention. The results are consistent with observations of STT model without Attention. The model configuration ‘MSCOCO-FLICKR 30K’ indicates that the model is trained on MSCOCO [7] and evaluated on FLICKR 30K [27]. Tables 26 and 27 indicate that transfer learning from large scale datasets like MSCOCO [7] to small scale datasets like FLICKR 30K [27] performs better.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis work presents a unified model which can generalize well on a diverse set of tasks namely image captioning, cross modal retrieval and sentence paraphrasing. This work shows that joint training of the models on various tasks enforces a tight interplay between vision and language. This model emulates the human brain which can perform multiple tasks simultaneously. In addition to the baseline STT model, attention mechanisms are introduced to align image regions and individual words in a sentence. The attention modules show a significant improvement in performance which indicates that the more complicated architectures learn better representations between different modalities.

5.2 Future work

This thesis work presents a simple architecture for aligning and generating new captions from images. Some of the possible extensions for this work are:

- Discovering new datapoints in this common vector space through sampling approaches and decoding them by using the sentence decoder.
- Adding image as a supervision to the sentence encoder while encoding the input sentence by passing the input image features to the initial timestep of the GRU network. This can be considered as an early fusion of images and captions.

- Incorporating text-text attention on the sentence paraphrases which can help the model learn the semantics between paraphrases.

Bibliography

[1] Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. "Multimodal deep learning." In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689-696. 2011.

[2] Reed, Scott, Zeynep Akata, Honglak Lee, and Bernt Schiele. "Learning deep representations of fine-grained visual descriptions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 49-58. 2016.

[3] M.-E. Nilsback and A. Zisserman, “Automated Flower Classification over a Large Number of Classes,” 2008, pp. 722–729.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” p. 8.

[5] R. Gandhi, “Build Your Own Convolution Neural Network in 5 mins,” Towards Data Science, 18-May-2018. [Online]. Available: https://towardsdatascience.com/build-your-own-convolution-neural-network-in-5-mins-4217c2cf964f. [Accessed: 18-Jul-2018].

[6] F. Rahman, seq2seq: Sequence to Sequence Learning with Keras. 2018.

[7] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” arXiv:1405.0312 [cs], May 2014.

[8] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.

[9] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[10] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In CVPR, vol. 1, no. 2, p. 3. 2017.

[11] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115, no. 3 (2015): 211-252.

[12] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823. 2015.

[13] Wang, Jian, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. "Deep metric learning with angular loss." In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2612-2620. IEEE, 2017.

[14] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality Reduction by Learning an Invariant Mapping,” 2006, vol. 2, pp. 1735–1742.

[15] Oh Song, Hyun, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. "Deep metric learning via lifted structured feature embedding." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4004-4012. 2016.

[16] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.

[17] Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop (Vol. 2).

[18] Hoffer, Elad, and Nir Ailon. "Deep metric learning using triplet network." International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015.

[19] Duan, Yueqi, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. "Deep Adversarial Metric Learning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2780-2789. 2018.

[20] Olivier Moindrot blog, Link : https://omoindrot.github.io/triplet-loss

[21] Wu, Chao-Yuan, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. "Sampling matters in deep embedding learning." In Proc. IEEE International Conference on Computer Vision (ICCV). 2017.

[22] Eisenschtat, A., & Wolf, L. (2017, July). Linking image and text with 2-way nets. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[23] Vendrov, Ivan, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. "Order-embeddings of images and language." arXiv preprint arXiv:1511.06361 (2015).

[24] Faghri, Fartash, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives." (2017).

[25] Wehrmann, Jonatas and Barros Rodrigo. "Bidirectional Retrieval Made Simple" (2017).

[26] You, Quanzeng, Zhengyou Zhang, and Jiebo Luo. "End-to-End Convolutional Semantic Embeddings." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5735-5744. 2018.

[27] FLICKR30K dataset, Link: http://shannon.cs.illinois.edu/DenotationGraph/

[28] Huang, Yan, Qi Wu, and Liang Wang. "Learning semantic concepts and order for image and sentence matching." In Computer Vision and Pattern Recognition, CVPR. 2017.

[29] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.

[30] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.

[31] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137. 2015.

[32] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning." In CVPR, vol. 1, no. 2, p. 3. 2017.

[33] Xu, Jun, Tao Mei, Ting Yao, and Yong Rui. "Msr-vtt: A large video description dataset for bridging video and language." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288-5296. 2016.

[34] Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. "Youtube-8m: A large-scale video classification benchmark." arXiv preprint arXiv:1609.08675 (2016).

[35] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[36] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, pp. 91-99. 2015.

[37] Lee, Kuang-Huei, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. "Stacked Cross Attention for Image-Text Matching." arXiv preprint arXiv:1803.08024 (2018).

[38] Engilberge, Martin, Louis Chevallier, Patrick Pérez, and Matthieu Cord. "Finding beans in burgers: Deep semantic-visual embedding with localization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3984-3993. 2018.

[39] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

[40] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60, no. 2 (2004): 91-110.

[41] LeCun, Yann A., Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. "Efficient backprop." In Neural networks: Tricks of the trade, pp. 9-48. Springer, Berlin, Heidelberg, 2012.

[42] Understanding LSTMs, Blog link: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[43] Sah, Shagan, Ameya Shringi, Dheeraj Peri, John Hamilton, Andreas Savakis, and Ray Ptucha. "Multimodal Reconstruction Using Vector Representation." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3763-3767. IEEE, 2018.

[44] Dosovitskiy, Alexey, and Thomas Brox. "Generating images with perceptual similarity metrics based on deep networks." In Advances in Neural Information Processing Systems, pp. 658-666. 2016.